Decision Tree Example with Categorical Input Variables

Problem

The dataset shown below has categorical attributes (Outlook, Temperature, Humidity, Wind) and the target variable is whether or not to play the match (Play Match). Build a Decision Tree for the same.

Day	**Outlook**	**Temperature**	**Humidity**	**Wind**	**Play Match**
D_1	Sunny	Hot	High	Weak	No
D_2	Sunny	Hot	High	Strong	No
D_3	Overcast	Hot	High	Weak	Yes
D_4	Rainy	Mild	High	Weak	Yes
D_5	Rainy	Cool	Normal	Weak	Yes
D_6	Rainy	Cool	Normal	Strong	No
D_7	Overcast	Cool	Normal	Strong	Yes
D_8	Sunny	Mild	High	Weak	No
D_9	Sunny	Cool	Normal	Weak	Yes
D_{10}	Rainy	Mild	Normal	Weak	Yes
D_{11}	Sunny	Mild	Normal	Strong	Yes
D_{12}	Overcast	Mild	High	Strong	Yes
D_{13}	Overcast	Hot	Normal	Weak	Yes
D_{14}	Rainy	Mild	High	Strong	No

Solution

Entropy of the Entire Dataset

To calculate the entropy of the entire dataset, we use the formula:

$$E = \sum_i -p_i \log p_i = -p_{\text{no}} \log_2(p_{\text{no}}) - p_{\text{yes}} \log_2(p_{\text{yes}})$$

where p_{yes} is the proportion of positive examples (Yes), and p_{no} is the proportion of negative examples (No).
In the dataset:

- Yes (Play Match) = 9 instances
- No (Play Match) = 5 instances
- Total = 14 instances

$$p_{\text{yes}} = \frac{9}{14}, \quad p_{\text{no}} = \frac{5}{14}$$

Now, we can plug these values into the entropy formula:

$$E = -\left(\frac{5}{14}\log_2\frac{5}{14}\right) - \left(\frac{9}{14}\log_2\frac{9}{14}\right)$$

$$E = 0.94$$

All possible splits

The splitting can happen either at the 'Weather' attribute or the 'Temperature' attribute or the 'Humidity' attribute or the 'Wind' attribute.

We will calculate the entropy and information gain for all the attributes and determine the splits for the decision tree.

Information Gain for the Attribute "Outlook"

We now calculate the entropy for each value of the Outlook attribute. There are three possible values for Outlook: Sunny, Overcast, and Rainy.

Entropy calculation for Sunny There are five days (namely D_1, D_2, D_8, D_9 and D_{11}) for which the Outlook is Sunny. Among these five days, three days (namely D_1, D_2 and D_8) have the output label (Match played) as 'No'. The other two days (namely D_8 and D_{11}) have the output (Match played) label as 'Yes'. Therefore, we have $p_{\text{no}} = \frac{2}{5}$ and $p_{\text{yes}} = \frac{3}{5}$. Hence, the entropy would be

$$E_{\text{sunny}} = -p_{\text{no}} \log_2(p_{\text{no}}) - p_{\text{yes}} \log_2(p_{\text{yes}}) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right)$$

Since $p_{\text{yes}} = 0$ and $p_{\text{no}} = 1$, the entropy becomes:

$$\Rightarrow E_{\text{sunny}} = 0.971$$

Entropy calculation for Overcast and Rainy Likewise, for Overcast and Rainy, the entropy values can be computed as

$$E_{\text{overcast}} = -\left(\frac{0}{4}\log_2\frac{0}{4} + \frac{4}{4}\log_2\frac{4}{4}\right) = 0$$

$$E_{\text{rainy}} = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) = 0.971$$

Weighted Average Entropy for Outlook

Out of the 14 days, since there are five days with sunny outlook, four days with overcast outlook and five days with rainy outlook, the corresponding weights would be $\frac{5}{14}$, $\frac{4}{14}$ and $\frac{5}{14}$, respectively. Hence, the weighted average of entropy for outlook is given by

$$E' = \frac{5}{14} \times E(\text{sunny}) + \frac{4}{14} \times E(\text{overcast}) + \frac{5}{14} \times E(\text{rainy})$$

$$\Rightarrow E' = \frac{5}{14} \times 0.971 + \frac{4}{14} \times 0 + \frac{5}{14} \times 0.971 = 0.694$$

Information Gain for Outlook

$$IG_{\text{outlook}} = E - E' = 0.94 - 0.694 = 0.247$$

Similarly, the information gain (IG) values for other attributes are calculated and shown in the table below.

Feature	i	Entropy (E_i)	Weighted Entropy (E')	Information Gain (IG)
Outlook	Sunny	$-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}=0.971$	$\frac{5}{14}E_{\text{sunny}}+\frac{4}{14}E_{\text{overcast}}+\frac{5}{14}E_{\text{rainy}}=0.694$	$E-E'=0.94-0.694=0.247$
	Overcast	$-\frac{0}{4}\log_2\frac{0}{4}-\frac{4}{4}\log_2\frac{4}{4}=0$		
	Rainy	$-\frac{2}{5}\log_2\frac{2}{5}-\frac{3}{5}\log_2\frac{3}{5}=0.971$		
Temperature	Hot	$-\frac{2}{4}\log_2\frac{2}{4}-\frac{2}{4}\log_2\frac{2}{4}=1$	$\frac{5}{14}E_{\text{hot}}+\frac{4}{14}E_{\text{mild}}+\frac{5}{14}E_{\text{cool}}=0.911$	$E-E'=0.94-0.911=0.029$
	Mild	$-\frac{2}{6}\log_2\frac{2}{6}-\frac{4}{6}\log_2\frac{4}{6}=0.918$		
	Cool	$-\frac{1}{4}\log_2\frac{1}{4}-\frac{3}{4}\log_2\frac{3}{4}=0.811$		
Humidity	High	$-\frac{4}{7}\log_2\frac{4}{7}-\frac{3}{7}\log_2\frac{3}{7}=0.985$	$\frac{7}{14}E_{\text{high}}+\frac{7}{14}E_{\text{normal}}=0.788$	$E-E'=0.94-0.788=0.152$
	Normal	$-\frac{1}{7}\log_2\frac{1}{7}-\frac{6}{7}\log_2\frac{6}{7}=0.592$		
Wind	Strong	$-\frac{3}{6}\log_2\frac{3}{6}-\frac{3}{6}\log_2\frac{3}{6}=1$	$\frac{6}{14}E_{\text{strong}}+\frac{8}{14}E_{\text{weak}}=0.892$	$E-E'=0.94-0.892=0.048$
	Weak	$-\frac{2}{8}\log_2\frac{2}{8}-\frac{6}{8}\log_2\frac{6}{8}=0.811$		

Branches After Splitting on Outlook attribute

From the table, since the information gain for 'Outlook' attribute is the highest amongst all, we split on Outlook. Hence, the Outlook attribute is considered to perform branching, which is shown in Figure 1. We now consider the branches for each Outlook condition.

Branch 1: Outlook = Sunny

For this branch, the dataset is reduced to:

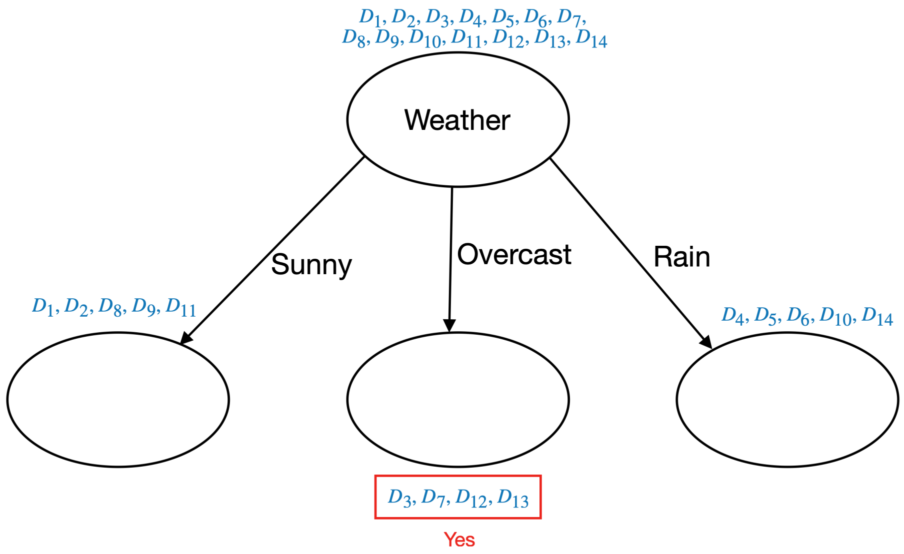
Day	**Outlook**	**Temperature**	**Humidity**	**Wind**	**Play Match**
D_1	Sunny	Hot	High	Weak	No
D_2	Sunny	Hot	High	Strong	No
D_8	Sunny	Mild	High	Weak	No
D_9	Sunny	Cool	Normal	Weak	Yes
D_{11}	Sunny	Mild	Normal	Strong	Yes

Since the output labels include both "No" and "Yes", further splits are needed for this branch.

$$E=-p_{\text{no}}\log_2(p_{\text{no}})-p_{\text{yes}}\log_2(p_{\text{yes}})=-\left(\frac{3}{5}\log_2\frac{3}{5}+\frac{2}{5}\log_2\frac{2}{5}\right)$$

$$E=0.971$$

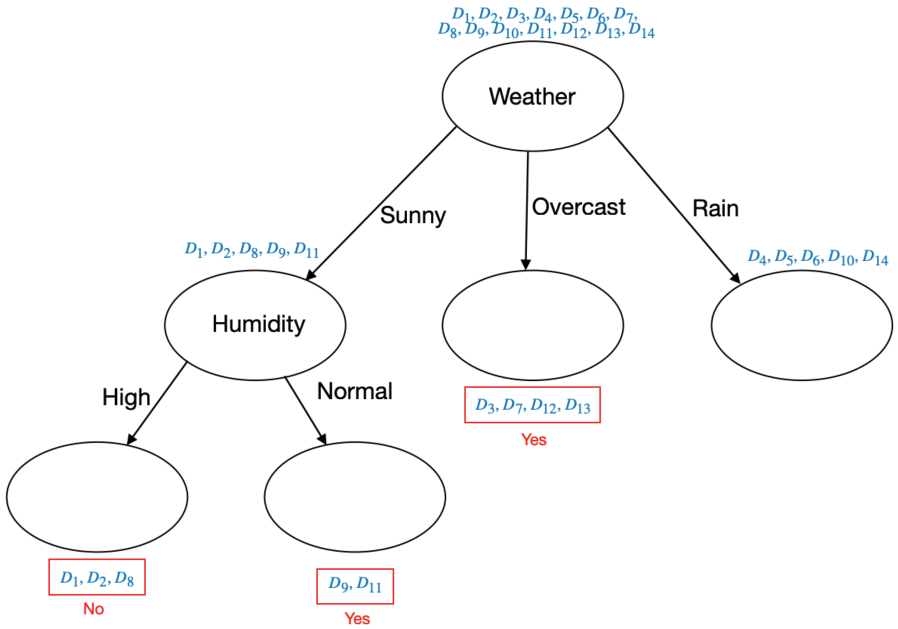
Figure 1: Branches after splitting on outlook attribute



Further splits Information Gain values are shown in the below table for all possible further splits. From the table, since the information gain for 'Humidity' attribute is the highest amongst all, we split on Humidity. Hence, the Humidity attribute is considered to perform branching, which is shown in Figure 2.

Feature	i	Entropy (E_i)	Weighted Entropy (E')	Information Gain (IG)
Temperature	Hot	$-\frac{2}{2}\log_2\frac{2}{2}-\frac{0}{2}\log_2\frac{0}{2}=0$	$\frac{2}{5}E_{\text{hot}}+\frac{2}{5}E_{\text{mild}}+\frac{1}{5}E_{\text{cool}}=0.4$	$E-E'=0.971-0.4=0.571$
	Mild	$-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}=1$		
	Cool	$-\frac{0}{1}\log_2\frac{0}{1}-\frac{1}{1}\log_2\frac{1}{1}=0$		
Humidity	High	$-\frac{3}{3}\log_2\frac{3}{3}-\frac{0}{3}\log_2\frac{0}{3}=0$	$\frac{3}{5}E_{\text{high}}+\frac{2}{5}E_{\text{normal}}=0$	$E-E'=0.971-0=0.971$
	Normal	$-\frac{0}{2}\log_2\frac{0}{2}-\frac{2}{2}\log_2\frac{2}{2}=0$		
Wind	Strong	$-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}=1$	$\frac{2}{5}E_{\text{strong}}+\frac{3}{5}E_{\text{weak}}=0.951$	$E-E'=0.971-0.951=0.02$
	Weak	$-\frac{2}{3}\log_2\frac{2}{3}-\frac{1}{3}\log_2\frac{1}{3}=0.918$		

Figure 2: Branches after further splitting on humidity attribute



Here, both the resulted nodes are leaf nodes[1]. Hence, there is no further splitting required.

Branch 2: Outlook = Overcast

For this branch, the dataset is reduced to:

[1] A leaf node is a node in which all the points from the dataset belong to a single class.

Day	Outlook	Temperature	Humidity	Wind	Play Match
D_3	Overcast	Hot	High	Weak	Yes
D_7	Overcast	Cool	Normal	Strong	Yes
D_{12}	Overcast	Mild	High	Strong	Yes
D_{13}	Overcast	Hot	Normal	Weak	Yes

$$E = 0$$

Since all instances lead to "Yes", no further splits are needed for this branch.

Branch 3: Outlook = Rainy

For this branch, the dataset is reduced to:

Day	Outlook	Temperature	Humidity	Wind	Play Match
D_4	Rainy	Mild	High	Weak	Yes
D_5	Rainy	Cool	Normal	Weak	Yes
D_6	Rainy	Cool	Normal	Strong	No
D_{10}	Rainy	Mild	Normal	Weak	Yes
D_{14}	Rainy	Mild	High	Strong	No

Since the output labels include both "No" and "Yes", further splits are needed for this branch.

$$E = -p_{\text{no}} \log_2 (p_{\text{no}}) - p_{\text{yes}} \log_2 (p_{\text{yes}}) = -\left(\frac{2}{5} \log_2 \frac{2}{5} + \frac{3}{5} \log_2 \frac{3}{5}\right)$$

$$E = 0.971$$

Further splits From the below table, since the information gain for 'Wind' attribute is the highest amongst all, we split on Wind. Hence, the Wind attribute is considered to perform branching, which is shown in Figure 3.

Here, both the resulted nodes are leaf nodes. Hence, there is no further splitting required.

Feature	i	Entropy (E_i)	Weighted Entropy (E')	Information Gain (IG)
Temperature	Mild	$-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3}=0.918$	$\frac{3}{5}E_{\text{mild}}+\frac{2}{5}E_{\text{cool}}=0.951$	$E-E'=0.971-0.951=0.02$
	Cool	$-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}=1$		
Humidity	High	$-\frac{1}{2}\log_2\frac{1}{2}-\frac{1}{2}\log_2\frac{1}{2}=1$	$\frac{2}{5}E_{\text{high}}+\frac{3}{5}E_{\text{normal}}=0.951$	$E-E'=0.971-0.951=0.02$
	Normal	$-\frac{1}{3}\log_2\frac{1}{3}-\frac{2}{3}\log_2\frac{2}{3}=0.918$		
Wind	Strong	$-\frac{2}{2}\log_2\frac{2}{2}-\frac{0}{2}\log_2\frac{0}{2}=0$	$\frac{2}{5}E_{\text{strong}}+\frac{3}{5}E_{\text{weak}}=0$	$E-E'=0.971-0=0.971$
	Weak	$-\frac{0}{3}\log_2\frac{0}{3}-\frac{3}{3}\log_2\frac{3}{3}=0$		

Figure 3: Branches after further splitting on wind attribute